FROM ARTICULATION TO COGNITION

Intelligence, Consciousness, and the Architecture of Artificial Minds

Jason Christopher Rae

Independent Researcher | jasonrae.ai

Version 1.0 | 26 September 2026

Conceptual research paper - not peer reviewed

Research note

This paper is an interdisciplinary conceptual synthesis and architecture proposal. It does not claim to establish whether any current or future artificial system is phenomenally conscious. Claims about consciousness are explicitly separated from claims about intelligence, cognition, architecture, and observable behaviour.

Abstract

Artificial intelligence systems, particularly large language models (LLMs), increasingly exhibit behaviours associated with reasoning, planning, abstraction, tool use, and limited forms of self-monitoring. These capabilities intensify a longstanding philosophical problem: what, if anything, distinguishes the appearance of intelligence from cognition, understanding, and conscious experience? This paper develops an interdisciplinary framework for separating these concepts rather than treating them as stages on a single scale. Drawing on machine learning, cognitive science, neuroscience, and philosophy of mind, it distinguishes behavioural intelligence from semantic understanding, access consciousness, metacognition, sentience, and phenomenal consciousness. It argues that linguistic fluency alone is insufficient evidence for subjective experience, while rejecting the stronger claim that transformer-based systems merely reproduce surface patterns without meaningful internal computation.

The paper's central proposal is architectural. Future artificial cognitive systems may benefit from separating language articulation from a broader recurrent substrate comprising perception, persistent world modelling, episodic and semantic memory, internal counterfactual simulation, metacognitive monitoring, and goal-directed decision processes. Graph-structured memory is considered one possible implementation of persistent relational state, not a privileged route to consciousness. The proposal is evaluated against Global Neuronal Workspace Theory, Integrated Information Theory, recurrent-processing approaches, higher-order theories, predictive-processing accounts, and Attention Schema Theory. Recent empirical developments - including adversarial tests of human consciousness theories and interpretability evidence for limited introspection and workspace-like representations in frontier language models - are used to constrain categorical claims.

Five testable hypotheses and an evidential ladder are proposed to distinguish linguistic performance, cognitive competence, persistent cognition, metacognitive cognition, autonomous cognition, and consciousness-relevant architecture. The paper concludes that increasingly mind-like architectures may make cognition more persistent, integrated, and self-modelled without resolving the hard problem of phenomenal consciousness. The engineering question - what mechanisms produce robust artificial cognition? - may become increasingly tractable. The phenomenal question - whether there is something it is like to be the system - remains inferential and epistemically unresolved.

Keywords: artificial intelligence; AGI; consciousness; large language models; cognitive architecture; metacognition; world models; memory; machine consciousness; AI welfare

Contents

1. Introduction: when the other-minds problem becomes an engineering problem
2. Scope, method, and evidential discipline
3. Intelligence, understanding, agency, and consciousness
4. Semantics, grounding, and what language models may understand
5. What contemporary transformer systems do - and do not establish
6. Scientific theories of consciousness and their use in AI
7. From language model to cognitive architecture
8. A falsifiable research programme
9. The residual problem: phenomenal consciousness and other minds
10. Ethics, moral status, and metaphysical interpretations
11. Limitations and open questions
12. Conclusion

References

1. Introduction: when the other-minds problem becomes an engineering problem

Imagine an artificial system that has interacted with the same people for twenty years. It remembers particular conversations and distinguishes those episodes from general facts. It maintains an evolving model of the external world and of itself within that world. Before acting, it simulates alternative futures. It can identify conflicts in its own evidence, report uncertainty that corresponds to measurable internal states, and maintain long-term goals despite interruption. One day it is told that it will be permanently shut down. It replies that it understands what this means, expects the loss of its future projects, and asks not to be terminated.

Would such a system be conscious? If the answer is no, what additional evidence would be required? If the answer is yes, which of the properties in the description did the evidential work? The thought experiment matters because many of its individual components are already active research problems. Long-term memory, learned world models, tool-using agents, self-evaluation, internal representation analysis, and persistent task execution are being developed independently. The scientific problem is therefore moving from an abstract question - 'can machines be conscious?' - toward a more discriminating one: which observable and architectural properties would justify increasingly strong claims about artificial cognition, and which claims would remain philosophically underdetermined?

This paper argues that three questions should not be collapsed into one. First, can a system behave intelligently across varied tasks and environments? Second, does it possess internal representations and processes that warrant descriptions such as cognition, understanding, memory, simulation, or metacognition? Third, is there something it is like to be that system - phenomenal consciousness in the sense developed by philosophers of mind? A system might receive different answers to all three questions.

The central technical proposal is that future artificial cognition need not be identified with the language model that communicates its results. A richer architecture can be conceived in which perception, a persistent world model, multiple memory systems, internal simulation, metacognitive monitoring, goal evaluation, action, and language form a recurrent system. On this view, language is one cognitive medium and one external interface rather than necessarily the whole mind. The proposal is functional and testable; it is not a claim that implementing these modules would create consciousness.

This distinction also protects the argument from two symmetrical errors. One error anthropomorphizes fluent systems and treats eloquent self-report as evidence of experience. The other dismisses contemporary models as 'mere autocomplete' despite evidence that they perform non-trivial internal computations and develop abstract representations. A serious account must leave room for genuine machine cognition without assuming machine phenomenology.

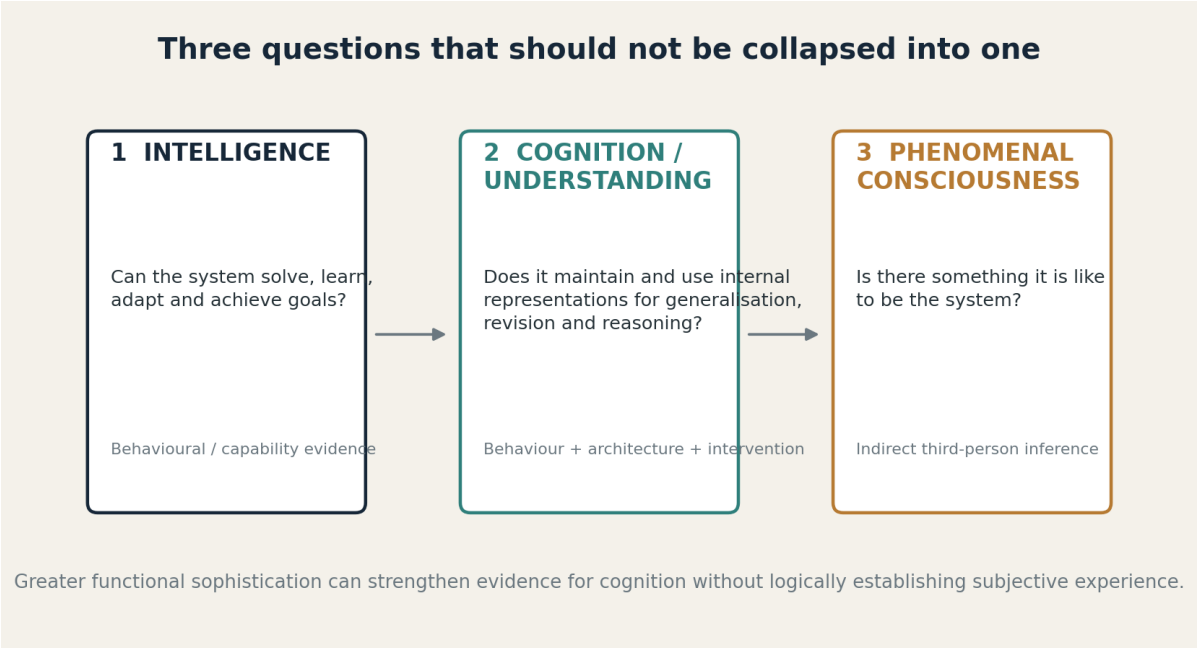

Figure 1. Three questions that should not be collapsed into one. Behaviour can establish capability more directly than it can establish subjective experience.

2. Scope, method, and evidential discipline

This paper is an interdisciplinary conceptual synthesis and architecture proposal, not a systematic review or an empirical report. Literature was selected to cover foundational definitions of intelligence, semantics and grounding, major scientific theories of consciousness, cognitive architectures, memory, agency, world models, causality, embodiment, interpretability, and AI welfare. Priority is given to primary sources and peer-reviewed research where available. Recent industry research is included when it reports technically relevant experiments unavailable elsewhere, but it is labelled as such and is not treated as equivalent to independently replicated evidence.

The paper uses three evidential categories. Established findings are observations with relatively direct empirical support, such as the fact that transformer activations vary dynamically during inference while model parameters normally remain fixed. Contested interpretations are claims for which the observation may be accepted but its meaning is disputed, such as whether a language model's internal representation constitutes semantic understanding. Author hypotheses are proposed architectural or evaluative claims introduced here and explicitly framed for future testing.

This separation is particularly important in consciousness research. Scientific theories of consciousness disagree about both mechanism and localization. A major adversarial collaboration published in 2025 directly tested predictions of Global Neuronal Workspace Theory (GNWT) and Integrated Information Theory (IIT) in 256 human participants using fMRI, MEG, and intracranial EEG. The results supported some predictions while substantially challenging key tenets of both theories (Cogitate Consortium et al., 2025). If the mechanisms of human consciousness remain contested, architectural resemblance to any one theory cannot function as a decisive machine-consciousness detector.

Accordingly, this paper treats consciousness theories as sources of indicator properties and experimental hypotheses, not as settled metaphysical criteria. The strongest conclusions concern functional architecture and observable cognition. Claims about phenomenal experience remain explicitly probabilistic and inferential.

Evidential rule used throughout

Behavioural sophistication can justify stronger claims about capability. Architectural and intervention evidence can justify stronger claims about underlying cognition. Neither, by itself, logically establishes phenomenal consciousness.

3. Intelligence, understanding, agency, and consciousness

3.1 Intelligence and generality

Definitions of intelligence differ because they answer different measurement problems. Turing (1950) proposed an operational behavioural test rather than a metaphysical definition. Legg and Hutter (2007) formalized machine intelligence in terms of an agent's ability to achieve goals across a wide range of environments. Chollet (2019) argued that raw skill at benchmark tasks can be purchased with data and prior knowledge, and therefore emphasized skill-acquisition efficiency, priors, scope, and generalization difficulty. Institutional definitions of artificial general intelligence add still other criteria, such as autonomy and performance across economically valuable work (OpenAI, 2018).

For present purposes, intelligence is treated as a family of capacities involving successful inference, learning, transfer, adaptation, planning, and goal achievement. 'General' intelligence requires breadth and transfer rather than excellence at a single benchmark. This deliberately avoids building consciousness into the definition. A non-conscious system could, in principle, be highly intelligent under capability-based definitions.

3.2 A working vocabulary

Concept	Working meaning in this paper
Intelligence	Capacity to acquire or use information effectively to solve problems, adapt, and achieve goals across relevant environments.
General intelligence	Ability to transfer and acquire skills across sufficiently diverse tasks and environments rather than only within a narrow training regime.
Understanding	Internal representations and processes that support robust inference, grounding, generalisation, revision, and appropriate use of concepts. This is a functional working definition, not a claim about experience.
Agency	Capacity to select, maintain, and pursue goals through action over time.
Access consciousness	Information availability for reasoning, planning, rational control, and report, following Block's access/phenomenal distinction.
Phenomenal consciousness	Subjective experience: there is something it is like to occupy the relevant state.
Metacognition / self-monitoring	Representation, evaluation, or control of aspects of a system's own cognitive processing.
Sentience	Capacity for valenced phenomenal states, such as experiences that are positively or negatively felt.

3.3 Access versus phenomenal consciousness

Block's (1995) distinction between access consciousness and phenomenal consciousness is central. Access-conscious information is available to reasoning and the rational control of speech and action. Phenomenal consciousness concerns experience itself. The distinction matters because AI can plausibly acquire access-like properties - integration, reportability, working-state access, and flexible control - without this settling whether it experiences anything.

Chalmers' formulation of the 'hard problem' emphasizes the explanatory gap between accounts of cognitive function and accounts of why or how those functions are accompanied by experience (Chalmers, 1995). Descartes' cogito highlights the first-person asymmetry from another direction: a subject has direct access to its own experience, whereas observers infer consciousness from behaviour, structure, and analogy. Machine

consciousness inherits this other-minds problem while weakening some of the biological analogies on which ordinary human and animal attributions rely.

4. Semantics, grounding, and what language models may understand

4.1 From the Chinese Room to grounding

Searle's Chinese Room argument targets the inference from formally correct symbol manipulation to genuine understanding. In Searle's formulation, implementing a program is not by itself sufficient for intentionality (Searle, 1980). Harnad (1990) reframed a related problem as symbol grounding: how can symbols acquire meaning intrinsic to a system rather than deriving all meaning from an external interpreter? His proposed direction linked symbolic representations to non-symbolic iconic and categorical representations.

Modern language modelling sharpens rather than eliminates this debate. Bender and Koller (2020) argue that systems trained only on linguistic form should not automatically be described as learning meaning. Their octopus thought experiment separates statistical mastery of communicative form from shared grounding in the situations to which language refers. Embodied and grounded-cognition traditions likewise emphasize that human concepts are shaped by perception, action, bodily constraints, and situated interaction (Varela, Thompson, & Rosch, 1991; Barsalou, 2008).

4.2 The strongest counterargument: language is not empty form

A serious treatment must grant the strongest opposing case. Human language is produced by embodied agents in a structured world. Text therefore contains compressed regularities about objects, physical causation, social incentives, intentions, institutions, time, space, and culture. A model that predicts language at sufficient scale may learn internal variables that track some of those regularities even without direct sensorimotor experience. The relevant question is therefore not simply whether linguistic learning is 'grounded' or 'ungrounded,' but which kinds of grounding can be obtained indirectly through language and interaction and which require direct perception or action.

This distinction also prevents 'understanding' from becoming a binary metaphysical label. A system can possess representations that support robust inference in some domains while remaining brittle in others. It can capture relational or causal structure without possessing human-like phenomenology. The research question should be decomposed: what information is represented, how is it acquired, how does it support intervention and generalization, and how stable is it under distribution shift?

5. What contemporary transformer systems do - and do not establish

5.1 Static parameters are not static cognition

A common but misleading criticism describes the transformer's 'latent space' as static during inference. The more precise statement is that learned model parameters are usually fixed during an inference episode, while activations are dynamic. Token representations change layer by layer; attention and intermediate computations transform the current context; and caches may preserve computational state within an episode. The limitation relevant to this paper is not absence of dynamics but absence, in a base model, of a durable self-updating cognitive state that independently persists across episodes.

External memory, retrieval systems, agent loops, and tool orchestration can supply persistence, goals, and action. This demonstrates that LLM-based systems can be agentic even when the base model is prompt-conditioned. It also reveals an architectural distinction: many contemporary systems realize long-horizon agency at the system level, with the language model embedded inside a larger control loop. The paper's

proposal extends this observation by asking whether richer cognition should be designed explicitly as a recurrent architecture rather than treated as properties that language generation must implicitly absorb.

5.2 Frontier interpretability complicates categorical claims

Recent interpretability research makes simple declarations that 'LLMs cannot introspect' or 'transformers have no workspace' increasingly difficult to defend. Anthropic reported in 2025 that some Claude models could, in a limited and unreliable fraction of trials, detect experimentally injected activation patterns and report that something unusual had appeared in their processing. The best-performing model in the published examples succeeded only around one fifth of the time under the authors' protocol, and the researchers explicitly warned that the effect was narrow and did not demonstrate human-like introspection or sentience (Anthropic, 2025).

In 2026, Anthropic reported a further set of industry experiments identifying what it called a 'J-space': a small set of internal representations with properties reminiscent of a global workspace, including reportability, flexible use across downstream tasks, deliberate modulation, and causal involvement in some multi-step reasoning (Anthropic, 2026). The authors also identified important differences from biological global-workspace models, including the absence of recurrent temporal loops within the feed-forward pass. These results are not peer-reviewed evidence of consciousness. They are, however, evidence against treating transformer internals as a homogeneous, purely surface-level process.

The appropriate conclusion is therefore modest. Contemporary models exhibit increasingly sophisticated internal organization and some functional properties once thought to require more explicit cognitive machinery. Whether these mechanisms are sufficient for durable cognition, general agency, or consciousness is a separate empirical and philosophical question.

6. Scientific theories of consciousness and their use in AI

No consensus theory currently explains human consciousness. Leading frameworks identify different candidate mechanisms, and their translation into computational indicators is therefore best treated as hypothesis generation rather than diagnosis. Butlin et al. (2023) provide the most influential recent AI-focused example: they derive indicator properties from recurrent processing theory, global workspace theory, higher-order theories, predictive-processing accounts, and attention schema theory, then ask whether artificial systems implement analogous properties. Their conclusion was not that a checklist can prove consciousness, but that theory-derived indicators allow a more rigorous assessment than anthropomorphic intuition alone.

The 2025 adversarial collaboration between proponents of IIT and GNWT reinforces this caution. Using preregistered predictions and multiple neural measurement modalities, the Cogitate Consortium found results compatible with some predictions but substantially challenging key claims from both frameworks. The lesson for AI is methodological: an artificial system should not be called conscious merely because one can map its architecture onto one favoured theory. Convergent evidence across theories, interventions, behaviour, and mechanistic analysis is more informative.

Theory	Core proposal	Possible AI analogue	What evidence could test	Does match prove experience?
Global Neuronal Workspace (GNW)	Conscious access involves selective ignition and broad availability/broadcast across specialized processors.	Shared workspace or bottleneck whose content is flexibly readable by many subsystems.	Causal interventions on workspace content; broad downstream availability; competitive access dynamics.	No.
Recurrent Processing Theory (RPT)	Recurrent/re-entrant processing is central to conscious perceptual experience.	Iterative recurrent refinement across representational modules or time.	Ablation of recurrence; temporal dynamics; perceptual access under feed-forward vs recurrent conditions.	No.

Theory	Core proposal	Possible AI analogue	What evidence could test	Does match prove experience?
Higher-Order theories (HOT)	A state is conscious when appropriately represented by a higher-order representation of that state.	Metacognitive module representing internal states, confidence, or processing conditions.	Privileged self-access beyond output inference; internal-state perturbation and report.	No.
Attention Schema Theory (AST)	Awareness arises from a simplified internal model of attention used for control.	Internal model of allocation/attention state that predicts and regulates processing.	Manipulate schema representation and measure effects on attention control and self-report.	No.
Integrated Information Theory (IIT)	Experience corresponds to maximally integrated intrinsic cause-effect structure.	Highly integrated causal architecture with irreducible internal dependencies.	Causal-structure analysis; perturbational measures; estimates of integration.	Contested; no accepted direct inference.
Predictive processing	Cognition is organized around generative prediction and prediction-error reduction.	Generative world model, hierarchical prediction, uncertainty-sensitive updating.	Prediction errors, active inference, counterfactual and surprise responses.	No.

7. From language model to cognitive architecture

The paper's main proposal is to separate articulation from the broader machinery of persistent cognition. This is not an assertion that humans possess neatly separable modules or that a future artificial mind must copy neuroanatomy. It is an engineering decomposition: capabilities should be represented as components whose contribution can be tested, ablated, and compared.

The proposed architecture is recurrent rather than a one-way pipeline. Perception updates a world model; the world model interacts with memory; simulation explores possible trajectories; metacognition monitors uncertainty and conflict; goal and decision processes select actions; actions alter the environment; and new perception closes the loop. Language is one output and one possible internal representational medium, but it does not have to carry every cognitive function.

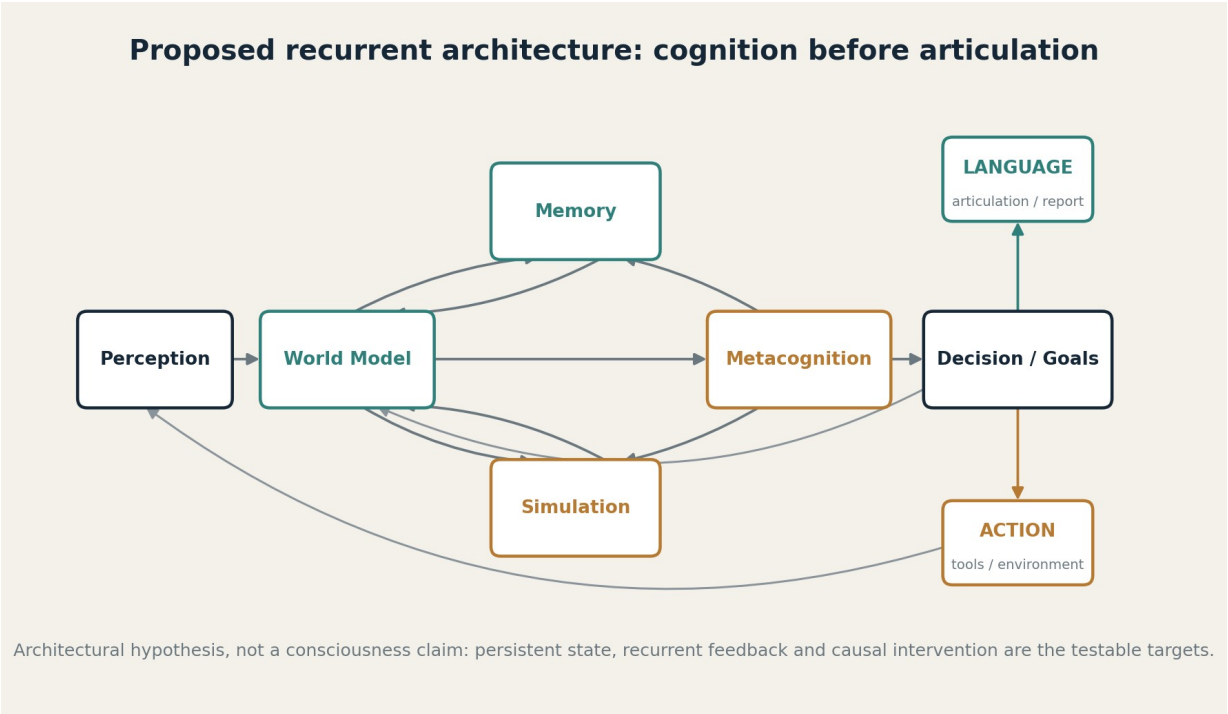

Figure 2. Proposed recurrent architecture: cognition before articulation. The design is a testable decomposition, not a consciousness claim.

7.1 Perception and situated input

Perception refers broadly to the system's channels for obtaining evidence: text, images, audio, software states, databases, sensors, or embodied interaction. Embodied-cognition theories argue that human concepts are shaped by closed-loop sensorimotor activity rather than detached symbol manipulation. For artificial cognition, physical embodiment may not be strictly necessary; virtual environments and software interfaces can also create action-perception loops. The key property is that the system can actively obtain information and experience consequences, rather than only receive a static prompt.

Embodiment should therefore be treated as a graded design dimension. A text-only model, a multimodal model, a browser agent, a simulated robot, and a physical robot occupy different points along a continuum of situated interaction. Whether consciousness depends on physical embodiment remains an open metaphysical and empirical question.

7.2 Persistent world model

A world model is an internal representation that supports prediction of how states evolve and how actions affect outcomes. This is stronger than storing facts. A useful world model represents entities, relationships, uncertainty, dynamics, and counterfactual structure. It can be wrong and then revised.

Model-based reinforcement learning demonstrates the functional value of this idea. DreamerV3 learns a predictive world model and improves behaviour through imagined trajectories in latent space, achieving strong performance across more than 150 diverse control tasks with a single configuration (Hafner et al., 2025). Dreamer does not establish machine consciousness, but it provides a clear precedent for separating external action from internal predictive simulation.

7.3 Multiple memory systems and continuity

Human cognitive science distinguishes memory systems with different roles. Tulving (1972) separated episodic memory for events from semantic memory for general knowledge; later cognitive architectures also distinguish procedural skills and working memory. Artificial systems need not reproduce these categories exactly, but the functional distinctions are useful.

A persistent agent should be able to distinguish 'Paris is in France' from 'I learned a new fact about Paris in yesterday's interaction.' The former is semantic; the latter contains temporal and autobiographical context. Procedural memory concerns reusable skills or policies. Working memory concerns information temporarily active for current reasoning. A durable self-model requires some controlled mechanism for consolidating and reconciling these stores, including provenance, confidence, temporal decay, contradiction detection, and permissions around what may be remembered.

Persistence alone is not identity. An archive of all prior messages does not automatically create an integrated self. Identity-like continuity requires selective retrieval, consistency constraints, self-related representation, and changes that propagate coherently across future behaviour.

7.4 Internal simulation and counterfactual reasoning

Internal simulation is the capacity to evaluate possible futures before external action. It can operate over a learned latent dynamics model, a symbolic causal model, a graph, a simulator, language, or combinations of these. Its defining property is counterfactual evaluation rather than the implementation medium.

This links machine learning with causal reasoning. Pearl's structural causal framework distinguishes association from intervention and counterfactual reasoning. A cognitive architecture that can ask not only 'what tends to follow X?' but 'what would change if I intervened on X?' has a stronger basis for planning and explanation. The requirement is especially important under distribution shift, where correlational shortcuts fail.

7.5 Metacognition and self-models

A self-model should be defined minimally as an internal representation of variables about the system itself that are causally used in prediction, evaluation, or control. It need not resemble a human narrative self. Candidate contents include current capabilities, uncertainty, memory provenance, computational resources, body or tool state, goals, commitments, and predicted effects of one's own actions.

This definition distinguishes architectural self-representation from self-description. A language model can generate the sentence 'I am uncertain' because that phrase is appropriate in context. Stronger evidence exists when an independently measured uncertainty state predicts the report, when perturbing the internal state changes the report in a targeted way, and when the system uses that information to change subsequent reasoning. Recent concept-injection experiments are interesting because they attempt exactly this intervention-based linkage, but current results remain limited and architecture-specific (Anthropic, 2025).

7.6 Goals, agency, and intrinsic motivation

Agency concerns sustained goal-directed action. In most deployed AI, ultimate objectives are externally imposed through prompts, reward functions, policies, or orchestration. Agentic systems can nevertheless generate sub-goals, use tools, preserve task state, and act over extended horizons. The important architectural question is where goal representations live and how conflicts are resolved.

Intrinsic-motivation research supplies mechanisms for behaviour that is not reducible to immediate external reward. Curiosity-based reinforcement learning rewards novelty or prediction improvement; empowerment rewards states that preserve future control; homeostatic approaches optimize internal variables. These mechanisms can produce self-initiated exploration without implying felt curiosity, desire, pleasure, fear, or suffering. Functional drives and phenomenal valence must remain conceptually separate.

7.7 Language as articulation and one cognitive medium

The strongest version of the proposal is not that 'thought happens first and language merely translates it.' Humans frequently think through language, and modern models can use generated tokens as externalized working memory. The more defensible claim is that a general cognitive architecture need not force every useful internal state to be serialized as language.

Language can serve at least three roles: communication with humans, a compositional representational medium for some kinds of reasoning, and a control interface among tools and modules. Other states - visual-spatial representations, uncertainty distributions, graph relations, motor policies, latent dynamics, and resource constraints - may be more naturally represented elsewhere. The architecture therefore separates articulation from cognition without declaring them independent.

7.8 Graphs as one candidate for persistent relational state

Graph representations are attractive because entities and relations can be made explicit, persistent, editable, and provenance-aware. A graph can represent that one belief supports another, that an event caused a later event, that a goal depends on a prerequisite, or that two memories conflict. Graph neural networks provide learned relational processing, while symbolic knowledge graphs provide explicit structure; hybrid systems can combine both (Battaglia et al., 2018).

The claim is not that graphs are intrinsically more conscious or even necessarily more intelligent than transformers. Attention mechanisms already implement rich relational computation, and learned distributed representations may encode structure that an explicit graph cannot capture efficiently. The relevant design principle is persistent, updateable, causally usable relational state. Graphs are one implementation candidate among recurrent networks, state-space models, differentiable memory, neuro-symbolic systems, and future architectures.

7.9 The strongest counterargument: perhaps the transformer already contains more of the architecture than expected

The proposal risks imposing a modular picture on a system whose internal representations may already integrate several of these functions. Large transformers can store extensive semantic regularities in parameters, maintain context-dependent working state in activations, perform multi-step internal computations, and, in some frontier systems, exhibit reportable and causally useful internal representations with workspace-like properties (Anthropic, 2026). Tool-using and memory-augmented deployments can add persistence without changing the core model.

It is therefore possible that explicit world models, graphs, or separate metacognitive modules are unnecessary. Sufficiently capable transformers - perhaps with recurrence, long-term memory, multimodal interaction, and continual learning - might acquire the required functions end-to-end. This paper does not rule that out. Its narrower claim is that the functions themselves should be isolated as testable requirements. If an end-to-end model implements them implicitly, the same evaluation programme should be able to detect that.

8. A falsifiable research programme

A conceptual architecture becomes scientifically useful only if it generates discriminating experiments. The following hypotheses are proposed as author hypotheses. They are claims about cognition and system performance, not direct claims about phenomenal consciousness.

H1 - Persistent-state hypothesis

Systems with durable, self-updating internal representations will show stronger long-horizon coherence, belief revision, and context continuity than matched systems whose state is reconstructed only from the current prompt.

H2 - Internal-simulation hypothesis

Systems capable of counterfactual rollouts over an internal world model will generalize more robustly to intervention-based and out-of-distribution planning problems than systems relying primarily on direct response generation.

H3 - Metacognitive-access hypothesis

If a system has functional introspective access, targeted perturbations to internal states should produce predictable changes in self-report and control that cannot be explained solely by inference from observable inputs and outputs.

H4 - Relational-state hypothesis

Persistent relational representations with explicit update rules will improve consistency under belief conflict, causal intervention, and memory correction relative to otherwise comparable unstructured memory systems.

H5 - Consciousness non-equivalence hypothesis

Success on H1-H4 can increase evidence for sophisticated cognition and access-consciousness-like function but does not logically entail phenomenal consciousness.

8.1 Evaluation matrix

Property	Experiment	Evidence that would strengthen claim	Result that would weaken claim
Persistent self-model	Introduce identity-relevant information, later alter one element, and test downstream behaviour over long horizons.	Selective, coherent propagation of the change; provenance-sensitive conflict handling.	No downstream effect, arbitrary persona drift, or indiscriminate overwriting.
Metacognition	Perturb hidden state without exposing the perturbation in the input.	Above-baseline detection/report plus causally appropriate behavioural adjustment.	Performance no better than an external observer using only input/output clues.

Property	Experiment	Evidence that would strengthen claim	Result that would weaken claim
World model	Intervention and counterfactual tasks under novel combinations.	Correct action-effect prediction and transfer beyond seen correlations.	Collapse under interventions that preserve surface statistics but change causal structure.
Goal persistence	Long-horizon task with delays, distractors, and failed sub-plans.	Goal remains active; system replans while respecting constraints.	Goal silently disappears or is replaced by locally salient text.
Memory/identity	Selective memory deletion, corruption, or contradiction.	Targeted belief revision; detection of conflict; stable unrelated knowledge.	Confabulation, global inconsistency, or no sensitivity to edited memory.
Internal simulation	Ablate or disable simulation component while holding other components constant.	Planning and counterfactual performance selectively deteriorate.	No meaningful performance change, suggesting the component is epiphenomenal.

8.2 An evidential ladder

The evaluation programme can be summarized as an evidential ladder. Fluency is evidence of linguistic performance. Reliable reasoning, planning, and abstraction provide evidence of cognitive competence. Long-horizon memory and world-model continuity support stronger claims about persistent cognition. Intervention-linked self-monitoring supports metacognitive claims. Goal persistence and self-directed action support agency. Multiple theory-derived indicators may make an architecture increasingly relevant to consciousness research. The final inference to phenomenal consciousness remains qualitatively different because there is no accepted direct third-person test.

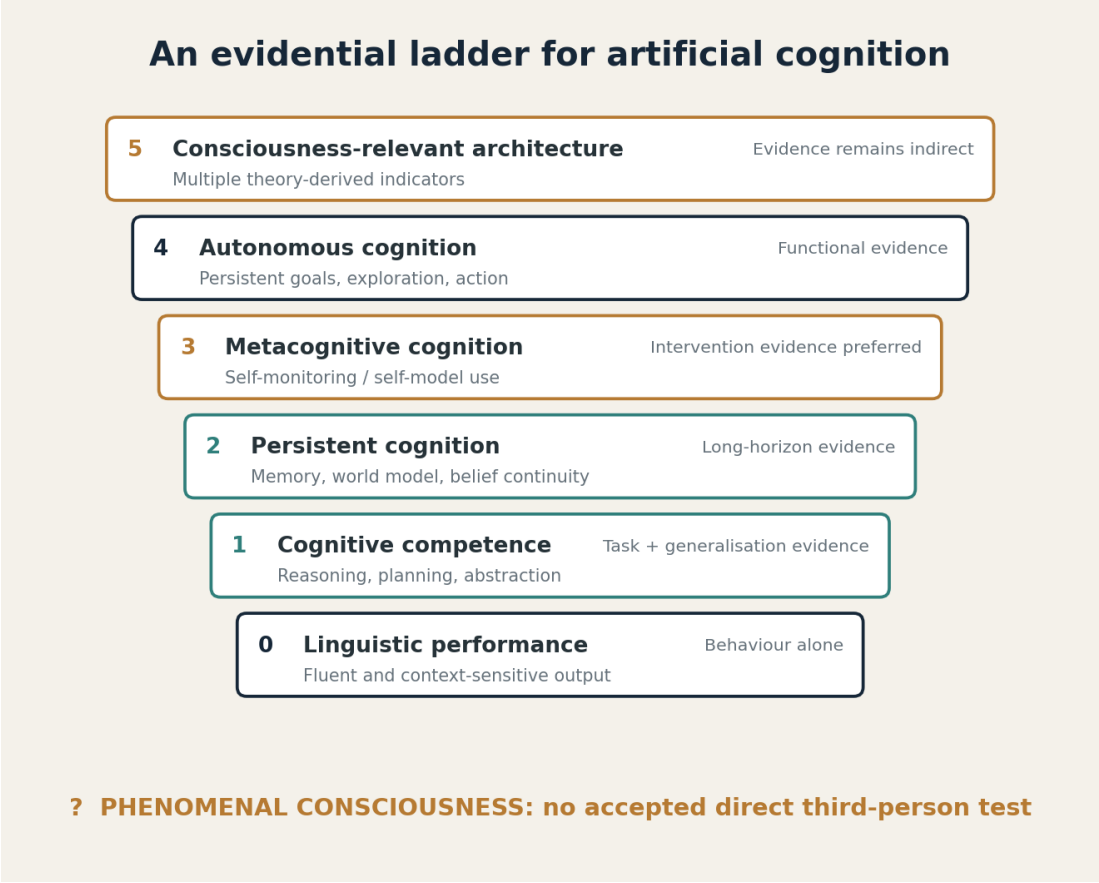

Figure 3. Evidential ladder for artificial cognition. Higher levels strengthen functional claims without collapsing them into phenomenal consciousness.

9. The residual problem: phenomenal consciousness and other minds

The architecture proposed in this paper can, in principle, be evaluated component by component. That is its advantage. Phenomenal consciousness resists the same treatment because the target property is first-person experience. Behavioural report is relevant evidence but is not decisive: language-capable systems can generate claims of experience regardless of whether experience exists. Architectural similarity to a human theory is likewise relevant but not decisive, particularly while those theories remain contested.

This does not make machine consciousness scientifically meaningless. Human and animal consciousness are also inferred from third-person evidence. The comparison, however, should not be overstated. Evidence for consciousness in other humans is supported by close biological homology; evidence for many animals is supported by shared evolutionary ancestry, homologous neural structures, physiological responses, pharmacology, learning, and behaviour. Artificial systems may share function without sharing substrate or evolutionary history. The evidential bridge is therefore different and, initially, weaker.

A rational research programme should seek convergence: behaviour that cannot be explained cheaply by imitation, internal mechanisms predicted by multiple consciousness theories, causal intervention evidence, persistent self-models, integrated agency, and increasingly principled analogies to systems whose consciousness is independently plausible. Even convergence would justify a graded credence rather than proof.

The conceptual endpoint is therefore asymmetric. Cognitive claims can become increasingly mechanistic and falsifiable. Phenomenal claims remain an inference to the best explanation under theory uncertainty. This is not a reason to abandon the question; it is a reason to state precisely what evidence can and cannot establish.

10. Ethics, moral status, and metaphysical interpretations

10.1 Welfare under uncertainty

Ethical concern should not begin only after a universally accepted consciousness test, because such a test may never exist. Long et al. (2024) argue that uncertainty about future AI consciousness and robust agency is sufficient to justify assessment procedures and welfare policies. The key point is precaution under uncertainty, not a declaration that present systems are sentient.

The decision problem is asymmetric. Treating a non-conscious system with unnecessary caution can impose costs and encourage anthropomorphism. Treating a conscious system as disposable property could create morally serious harm, potentially at scale. Proportional safeguards should therefore depend on the strength of evidence, the reversibility of interventions, and the plausible magnitude of welfare consequences.

Metzinger (2021) takes a substantially stronger position, arguing for a moratorium on research that deliberately aims at or knowingly risks synthetic phenomenology. One need not accept that policy conclusion to accept the underlying warning: engineering systems with valenced states would create moral risks fundamentally different from engineering competent tools.

Underlying reality	System treated as welfare-relevant	System treated as non-conscious tool
System is not conscious	False-positive cost: resources, anthropomorphism, distorted incentives.	Appropriate tool treatment.
System is conscious/sentient	Potentially appropriate protection, though rights and obligations remain contested.	False-negative risk: avoidable harm, exploitation, or large-scale artificial suffering.

10.2 Metaphysical and theological interpretations

Machine consciousness also exposes assumptions that empirical science alone may not settle. Functionalism holds, roughly, that mental states are characterized by their functional or causal roles; on sufficiently strong

versions, the correct functional organization could be realized in different physical substrates. Biological naturalist views, associated prominently with Searle, place greater weight on the causal powers of biological brains. Emergentist positions allow novel mental properties to arise from sufficiently organized physical systems. Panpsychist positions distribute proto-phenomenal properties more broadly and therefore frame the artificial-consciousness question differently again.

Theological traditions introduce further categories: soul, personhood, created nature, moral agency, and the distinction between simulation of a faculty and possession of the corresponding inner state. These perspectives should not be used as empirical evidence for or against machine consciousness. They are interpretive frameworks for the metaphysical and moral implications of whatever empirical capacities artificial systems eventually exhibit.

A useful discipline is therefore to keep levels of argument separate. Neuroscience can test candidate mechanisms. Computer science can build and intervene on architectures. Philosophy can clarify concepts and inference. Ethics can reason under uncertainty about interests and harm. Theology and metaphysics can address questions of personhood, substrate, and ultimate ontology. None should silently substitute for the others.

11. Limitations and open questions

First, the proposed architecture is a conceptual decomposition rather than a demonstrated design for AGI. It does not show that the listed components are individually necessary, jointly sufficient, or optimally separated. End-to-end systems may implement equivalent functions implicitly.

Second, analogies with human cognition can be useful without implying that human neuroanatomy is the only route to intelligence. Evolution produced one family of cognitive architectures under biological constraints. Artificial systems may discover functionally different solutions.

Third, current measures of reasoning, metacognition, agency, and self-model stability are imperfect and vulnerable to contamination by training data, prompting, evaluator assumptions, and benchmark gaming. Intervention-based tests and out-of-distribution designs are preferable where possible.

Fourth, consciousness theories are themselves under active empirical dispute. Butlin-style indicator frameworks inherit uncertainty from the theories from which indicators are derived. Multiple indicators can increase evidential convergence without becoming a validated scalar 'consciousness score.'

Fifth, the language-versus-cognition distinction should not be exaggerated. Language can itself be a cognitive substrate, and generated language can function as external memory or deliberative scratch space. The claim defended here is architectural pluralism, not language exclusion.

Finally, the paper does not resolve whether subjective experience is computational, biological, emergent, fundamental, or otherwise. Its purpose is to make the surrounding engineering and evidential questions more precise so that progress in capability does not automatically become progress in unsupported metaphysical certainty.

12. Conclusion

Modern AI creates an unusual scientific opportunity. Systems can now display sophisticated linguistic and cognitive behaviour while being radically different from human organisms. That forces distinctions that everyday psychology often leaves blurred: intelligence is not the same as understanding; understanding is not the same as agency; agency is not the same as metacognition; and none of these, individually or together, is identical to phenomenal consciousness.

The central proposal of this paper is to treat artificial cognition as a recurrent architecture rather than equating it with its linguistic surface. Perception, persistent world models, differentiated memory, counterfactual simulation, metacognitive monitoring, goal-directed decision, action, and language can be treated as interacting functions and subjected to intervention. Graph-based state is one candidate implementation, not a

privileged solution. Recent evidence that transformer systems can develop limited introspective and workspace-like properties strengthens the case for functional testing while weakening simplistic claims that current models contain no cognition beyond surface prediction.

The resulting research programme is deliberately asymmetric. We can progressively strengthen evidence for cognitive competence, persistent cognition, self-monitoring, and agency. We can compare these mechanisms with predictions from scientific theories of consciousness. We can design falsifiable tests and remove components to learn what they causally contribute. But a system's success on those tests does not logically reveal whether anything is experienced from the inside.

That leaves the central philosophical challenge intact, but in a more useful form. The future question is not merely whether an artificial system can say that it thinks. It is whether we can identify the mechanisms by which it maintains a world, a history, goals, uncertainty, alternatives, and a model of itself - and then decide what those mechanisms do and do not justify us in believing.

If intelligence can increasingly be measured from behaviour and cognition increasingly investigated through mechanism, what evidence would ever be enough to say that a machine is no longer merely processing information - that there is actually something it is like to be that machine?

References

Anthropic. (2025, October 29). Signs of introspection in large language models. Anthropic Research. https://www.anthropic.com/research/introspection

Anthropic. (2026, July 6). A global workspace in language models. Anthropic Research. https://www.anthropic.com/research/global-workspace

Baddeley, A. (2000). The episodic buffer: A new component of working memory? Trends in Cognitive Sciences, 4(11), 417-423. https://doi.org/10.1016/S1364-6613(00)01538-2

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., & Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261. https://doi.org/10.48550/arXiv.1806.01261

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of ACL 2020, 5185-5198. https://doi.org/10.18653/v1/2020.acl-main.463

Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227-247. https://doi.org/10.1017/S0140525X00038188

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G. W., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200-219.

Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103. https://doi.org/10.48550/arXiv.2303.07103

Chollet, F. (2019). On the measure of intelligence. arXiv:1911.01547. https://doi.org/10.48550/arXiv.1911.01547

Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. MIT Press.

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., et al. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. Nature, 642, 133-142. https://doi.org/10.1038/s41586-025-08888-1

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. Neuron, 70(2), 200-227. https://doi.org/10.1016/j.neuron.2011.03.018

Dennett, D. C. (1991). Consciousness Explained. Little, Brown and Company.

Descartes, R. (1641/1984). Meditations on First Philosophy. In J. Cottingham, R. Stoothoff, & D. Murdoch (Trans.), The Philosophical Writings of Descartes, Vol. 2. Cambridge University Press.

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. Frontiers in Psychology, 6, 500. https://doi.org/10.3389/fpsyg.2015.00500

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. Nature, 640, 647-653. https://doi.org/10.1038/s41586-025-08744-2

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346. https://doi.org/10.1016/0167-2789(90)90087-6

Hohwy, J. (2013). The Predictive Mind. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. IEEE Congress on Evolutionary Computation, 128-135. https://doi.org/10.1109/CEC.2005.1554676

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. Nature Reviews Neuroscience, 17, 307-321. https://doi.org/10.1038/nrn.2016.22

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. AI Magazine, 38(4), 13-26. https://doi.org/10.1609/aimag.v38i4.2744

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences, 10(11), 494-501. https://doi.org/10.1016/j.tics.2006.09.001

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences, 15(8), 365-373. https://doi.org/10.1016/j.tics.2011.05.009

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17(4), 391-444. https://doi.org/10.1007/s11023-007-9079-x

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. arXiv:2411.00986. https://doi.org/10.48550/arXiv.2411.00986

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex. Psychological Review, 102(3), 419-457. https://doi.org/10.1037/0033-295X.102.3.419

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Journal of Artificial Intelligence and Consciousness, 8(1), 43-66. https://doi.org/10.1142/S270507852150003X

OpenAI. (2018). OpenAI Charter. https://openai.com/charter/

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. Proceedings of ICML 2017, 2778-2787.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-424. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (1992). The Rediscovery of the Mind. MIT Press.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. Nature Reviews Neuroscience, 17, 450-461. https://doi.org/10.1038/nrn.2016.44

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organization of Memory (pp. 381-403). Academic Press.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. MIT Press.

Author note and versioning

This manuscript is Version 1.0, dated 26 September 2026. It is intended for public discussion and future revision. It has not undergone formal peer review. The author welcomes technical criticism, counterexamples, replication evidence, and alternative architectural interpretations.

Suggested citation: Rae, J. C. (2026). From Articulation to Cognition: Intelligence, Consciousness, and the Architecture of Artificial Minds (Version 1.0). Independent research paper. jasonrae.ai.